Automatic discovery and processing of EEG cohorts from clinical records

TEMPLE UNIV OF THE COMMONWEALTH AND UNIVERSITY OF TEXAS - DALLAS

PI: PICONE, JOSEPH; HARABAGIU, SANDA MARIA; OBEID, IYAD   Grant Number: 1 U01 HG008468-01

Electronic medical records (EMRs) collected at every hospital in the country collectively contain a staggering wealth of biomedical knowledge. EMRs can include unstructured text, temporally constrained measurements (e.g., vital signs), multichannel signal data (e.g., EEGs), and image data (e.g., MRIs). This information could be transformative if properly harnessed. Information about patient medical problems, treatments, and clinical course is essential for conducting comparative effectiveness research. Uncovering clinical knowledge that enables comparative research is the primary goal of this proposal. We will focus on the automatic interpretation of clinical EEGs collected over 12 years at Temple University Hospital (over 25,000 sessions and 15,000 patients). Clinicians will be able to retrieve relevant EEG signals and EEG reports using standard queries (e.g. ""Young patients with focal cerebral dysfunction who were treated with Topamax""). In Aim 1 we will automatically annotate EEG events that contribute to a diagnosis. We will develop automated techniques to discover and time-align the underlying EEG events using semi-supervised learning. In Aim 2 we will process the text from the EEG reports using state-of-the-art clinical language processing techniques. Clinical concepts, their type, polarity and modality shall be discovered automatically, as well as spatial and temporal information. In addition, we shall extract the medical concepts describing the clinical picture of patients from the EEG reports. In Aim 3, we will develop a patient cohort retrieval system that will operate on the clinical knowledge extracted in Aims 1 and 2. In addition we shall organize this knowledge in a unified representation: the Qualified Medical Knowledge Graph (QMKG), which will be built using BigData solutions through MapReduce. The QMKG will be able to be searched by biomedical researchers as well as practicing clinicians. The QMKG will also provide a characterization of the way in which events in an EEG are narrated by physicians and the validation of these across a BigData resource. The EMKG represents an important contribution to basic science. In Aim 4 we will validate the usefulness of the patient cohort identification system by collecting feedback from clinicians and medical students who will participate in a rigorous evaluation protocol. Inclusion and exclusion criteria for the queries shall be designed and experts will provide relevance judgments for the results. For each query, medical experts shall examine the top-ranked cohorts for common precision errors (false positives) and the bottom five ranked common recall errors (false negatives). User validation testing will be performed using live clinical data and the feedback wil enhance the quality of the cohort identification system. The existence of an annotated BigData archive of EEGs will greatly increase accessibility for non- experts in neuroscience, bioengineering and medical informatics who would like to study EEG data. The creation of this resource through the development of efficient automated data wrangling techniques will demonstrate that a much wider range of BigData bioengineering applications are now tractable. PUBLIC HEALTH RELEVANCE PUBLIC HEALTH RELEVANCE: The primary goal of this proposal is to enable comparative research by automatically uncovering clinical knowledge from a vast BigData archive of

clinical EEG signals and EEG reports collected over the past 12 years at Temple University Hospital. In the proposed project, we will develop a proof-of-concept based on the discovery of patient cohorts and provide an annotated BigData archive as well as the software that enabled the annotations and the generation of the patient cohort retrieval system. This resource will be accompanied by a novel medical knowledge representation generated with MapReduce, greatly increasing accessibility for non- experts in neuroscience, bioengineering and medical informatics, and demonstrating the transformative potential of mining the staggering wealth of biomedical knowledge available in hospital medical records.